Fall 2014
Data-Intensive Systems

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

John Klein
Dr. Ian Gorton

October 29, 2014

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
29 OCT 2014	**N/A**	

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Data-Intensive Systems	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Klein /John	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Software Engineering Institute Carnegie Mellon University Pittsburgh, PA 15213	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited.

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**10**	

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM-0001797

Big Data Systems

NoSQL and horizontal scaling are changing architecture principles by creating *convergence of concerns*

- Can't abstract away underlying technology and topology - application, data, and deployment are tightly coupled
- Data technology selection is now an early decision, hard to change

Emerging big data databases complicate technology acquisition

Project Results:

- Technology selection method – rigorous, systematic, evidence-based
- Knowledge Base – decision support for architects and acquirers

Today's Warfighter has access to an ever-increasing number of sensors, imagers, internet artifacts, open source and other sophisticated collection devices, to the point that a major challenge has become how to sift through this massive amount of information to find the most critical and actionable items of intelligence.
'Big Data' tools, techniques, and technologies seek to provide the means to analyze, exploit and share conclusions drawn from this seemingly overwhelming information load.

Big Data Technology Evaluation Challenges

Rapidly changing technology landscape

- New products emerging, multiple releases per year on existing products
- Need to balance speed with precision

Large potential solution space

- Need to quickly narrow down and focus
- Products are *very* different – generalized comparisons are nearly impossible

Scale makes full-fidelity prototyping impractical

- Data sets, compute nodes, load generation

Technology is highly configurable

- Need to focus on go/no-go criteria
- Avoid trap of optimizing every test run

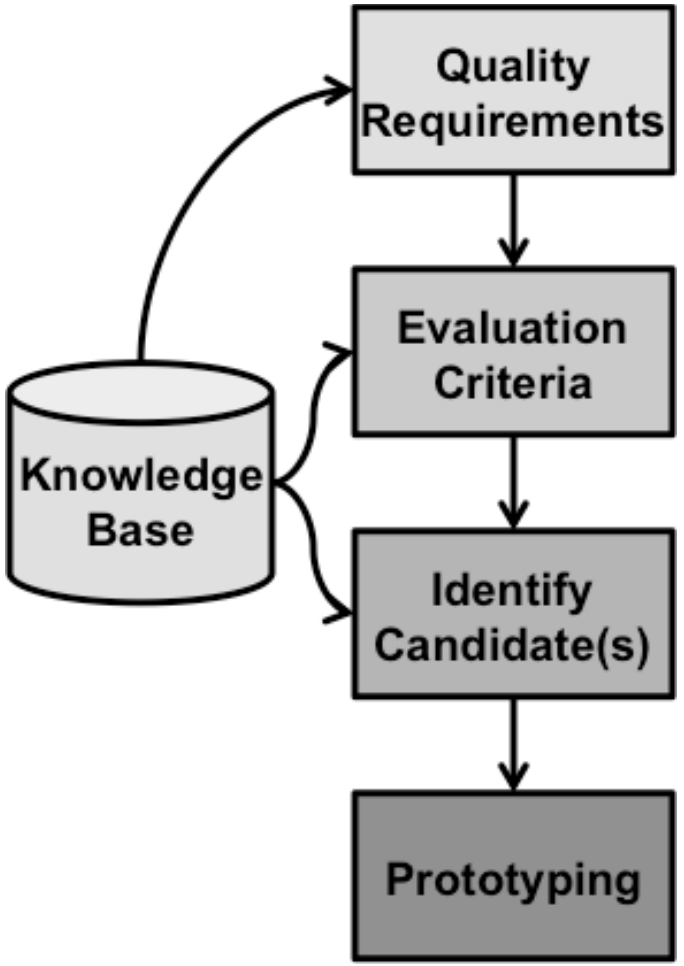
Lightweight Evaluation and Architecture Prototyping for Big Data (LEAP4BD)

Aims

- Risk reduction
- Rapid, streamlined selection/acquisition

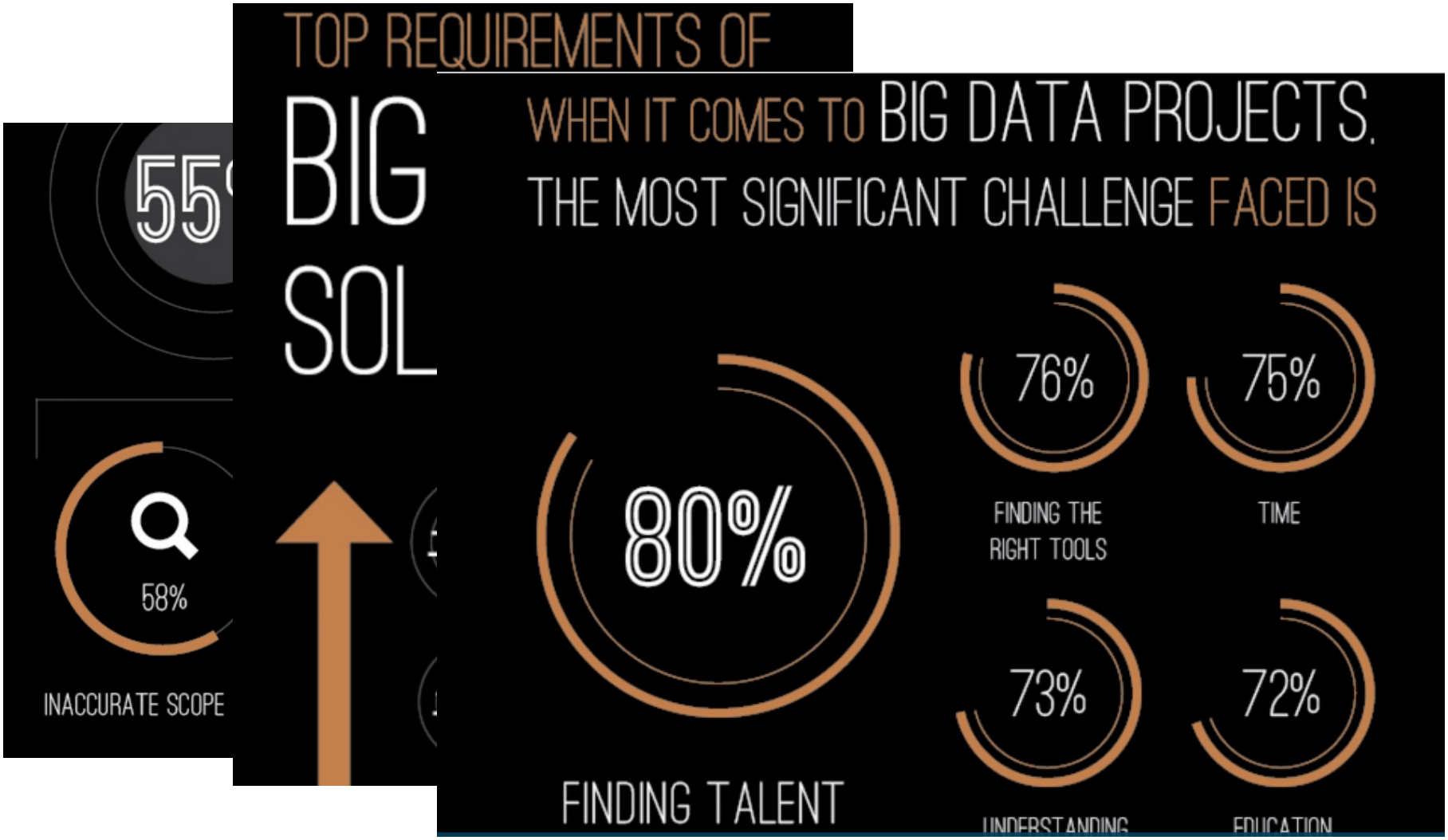
Steps

- Assess the system context and landscape
- Identify the architecturally-significant requirements and decision criteria
- Evaluate candidate technologies against quality attribute decision criteria
- Validate architecture decisions and technology selections through focused prototyping



Big Data Survey

http://visual.ly/cios-big-data



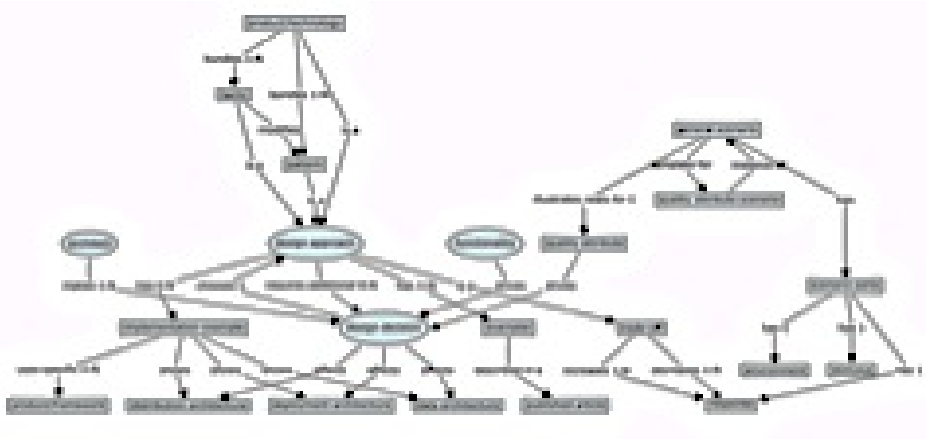
QuABase – A Knowledge Base for Big Data System Design

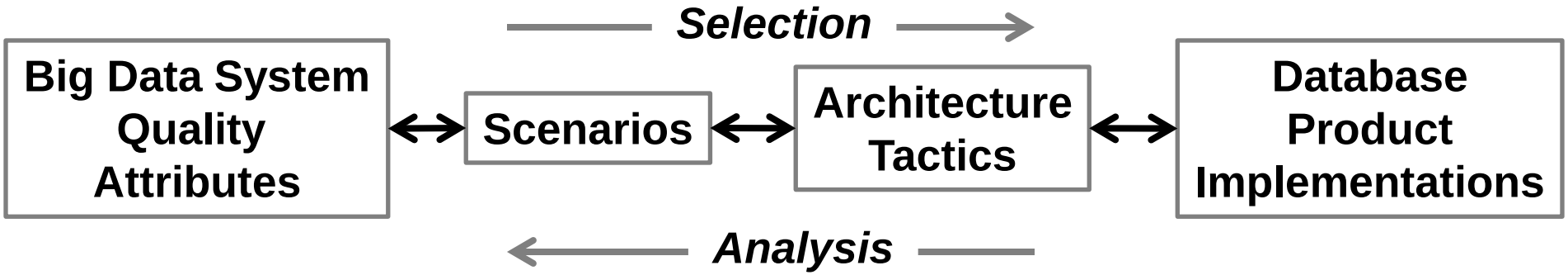
Semantic-based Knowledge Model

- General model of software architecture knowledge
- Populated with specific big data architecture knowledge

Dynamic, generated, and queryable content

Knowledge Visualization





QuABase – A Knowledge Base for Big Data System Design

Implemented on Semantic MediaWiki platform

- Knowledge schema is set of category triples
  - "*Database* implements *tactic*"
- Knowledge is set of instance triples
  - "*Cassandra* implements *multi-site replication*"
- Input forms embody and enforce schema
- Faceted and full-text search
- Graphical and tabular rendering of search results

Platform provides scalability and support for editing and curation workflows

Targeted to enable the "average" architect or acquirer to be successful working with big data infrastructures

Status

LEAP4BD

- Ready to pilot

QuABase

- Prototype is complete – covers 8 NoSQL/NewSQL implementations
- Completing validation testing

Big Data Architectures and Technologies

- 1 day instructor-led course with eLearning version in development

IEEE Software paper, SEI Blog posts

- ICSE paper in review

FY15 Research Plans

Apply machine learning to automate population of knowledge base

- Initial focus on NoSQL/NewSQL technology domain
- Extend to create knowledge bases in other key acquisition technology domains

Runtime observability for big data systems

- Efficiently define and deploy application-level monitors to assess system health after deployment
  - Systems run in shared environments with unpredictable QoS
  - User workloads evolve rapidly after deployment